

University of Dundee

Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration

Smid, Marcel; Rodríguez-González, F. Germán; Sieuwerts, Anieta M.; Salgado, Roberto; Prager-Van der Smissen, Wendy J. C.; Vlugt-Daane, Michelle van der

Published in:
Nature Communications

DOI:
[10.1038/ncomms12910](https://doi.org/10.1038/ncomms12910)

Publication date:
2016

Licence:
CC BY

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in Discovery Research Portal](#)

Citation for published version (APA):

Smid, M., Rodríguez-González, F. G., Sieuwerts, A. M., Salgado, R., Prager-Van der Smissen, W. J. C., Vlugt-Daane, M. V. D., van Galen, A., Nik-Zainal, S., Staaf, J., Brinkman, A. B., van de Vijver, M. J., Richardson, A. L., Fatima, A., Berentsen, K., Butler, A., Martin, S., Davies, H. R., Debets, R., Gelder, M. E. M-V., ... Martens, J. W. M. (2016). Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. *Nature Communications*, 7, 1-9. [12910]. <https://doi.org/10.1038/ncomms12910>

General rights

Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

ARTICLE

Received 15 Mar 2016 | Accepted 15 Aug 2016 | Published 26 Sep 2016

DOI: 10.1038/ncomms12910

OPEN

Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration

Marcel Smid^{1,*}, F. Germán Rodríguez-González^{1,*}, Anieta M. Sieuwerts¹, Roberto Salgado^{2,3}, Wendy J.C. Prager-Van der Smissen¹, Michelle van der Vlugt-Daane¹, Anne van Galen¹, Serena Nik-Zainal^{4,5}, Johan Staaf⁶, Arie B. Brinkman⁷, Marc J. van de Vijver⁸, Andrea L. Richardson^{9,10}, Aquila Fatima¹¹, Kim Berentsen⁷, Adam Butler⁴, Sancha Martin⁴, Helen R. Davies⁴, Reno Debets¹, Marion E. Meijer-Van Gelder¹, Carolien H.M. van Deurzen¹¹, Gaëtan MacGrogan¹², Gert G.G.M. Van den Eynden^{3,13}, Colin Purdie¹⁴, Alastair M. Thompson¹⁴, Carlos Caldas¹⁵, Paul N. Span^{16,17}, Peter T. Simpson¹⁸, Sunil R. Lakhani^{18,19}, Steven Van Laere²⁰, Christine Desmedt², Markus Ringnér⁶, Stefania Tommasi²¹, Jorunn Eyfjord²², Annegien Broeks²³, Anne Vincent-Salomon²⁴, P. Andrew Futreal²⁵, Stian Knappskog^{26,27}, Tari King²⁸, Gilles Thomas²⁹, Alain Viari^{29,30}, Anita Langerød^{31,32}, Anne-Lise Børresen-Dale^{31,32}, Ewan Birney³³, Hendrik G. Stunnenberg⁷, Mike Stratton⁴, John A. Foekens¹ & John W.M. Martens¹

A recent comprehensive whole genome analysis of a large breast cancer cohort was used to link known and novel drivers and substitution signatures to the transcriptome of 266 cases. Here, we validate that subtype-specific aberrations show concordant expression changes for, for example, *TP53*, *PIK3CA*, *PTEN*, *CCND1* and *CDH1*. We find that *CCND3* expression levels do not correlate with amplification, while increased *GATA3* expression in mutant *GATA3* cancers suggests *GATA3* is an oncogene. In luminal cases the total number of substitutions, irrespective of type, associates with cell cycle gene expression and adverse outcome, whereas the number of mutations of signatures 3 and 13 associates with immune-response specific gene expression, increased numbers of tumour-infiltrating lymphocytes and better outcome. Thus, while earlier reports imply that the sheer number of somatic aberrations could trigger an immune-response, our data suggests that substitutions of a particular type are more effective in doing so than others.

¹ Department of Medical Oncology, Erasmus MC Cancer Institute and Cancer Genomics Netherlands, Erasmus University Medical Center, 3015CN Rotterdam, The Netherlands.

² Breast Cancer Translational Research Laboratory, Université Libre de Bruxelles, Institut Jules Bordet, Bd de Waterloo 121, B-1000 Brussels, Belgium. ³ Department of Pathology/TCRU GZA, 2610 Antwerp, Belgium. ⁴ Wellcome Trust Sanger Institute, Hinxton CB10 1SA, Cambridge, UK. ⁵ East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 9NB, UK. ⁶ Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, SE-223 81 Lund, Sweden.

⁷ Faculty of Science, Department of Molecular Biology, Radboud Institute for Molecular Life Sciences, Radboud University Nijmegen, 6525GA, Nijmegen, The Netherlands.

⁸ Department of Pathology, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. ⁹ Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. ¹⁰ Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. ¹¹ Department of Pathology, Erasmus MC Cancer Institute, Erasmus University Medical Center, 3015CN Rotterdam, The Netherlands. ¹² Département de Biopathologie, Institut Bergonié, CS 61283 33076 Bordeaux, France. ¹³ Molecular Immunology Unit, Jules Bordet Institute, B-1000 Brussels, Belgium. ¹⁴ Department of Pathology, Ninewells Hospital & Medical School, Dundee DD1 9SY, UK. ¹⁵ Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. ¹⁶ Department of Radiation Oncology, Radboud University Medical Center, 6525GA, Nijmegen, The Netherlands. ¹⁷ Department of Laboratory Medicine, Radboud University Medical Center, 6525GA, Nijmegen, The Netherlands. ¹⁸ The University of Queensland: UQ Centre for Clinical Research and School of Medicine, Brisbane 4029, Australia. ¹⁹ Pathology Queensland, The Royal Brisbane and Women's Hospital, Brisbane 4029, Australia.

²⁰ Center for Oncological Research, University of Antwerp & GZA Hospitals Sint-Augustinus, 2610 Wilrijk, Belgium. ²¹ IRCCS Istituto Tumori 'Giovanni Paolo II', 70124 Bari, Italy.

²² Cancer Research Laboratory, Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland. ²³ The Netherlands Cancer Institute, 1066CX Amsterdam, The Netherlands.

²⁴ Department of Pathology and INSERM U934, Institut Curie, 26 rue d'Ulm, 75248 Paris Cedex 05, France. ²⁵ Department of Genomic Medicine, UT MD Anderson Cancer Center, Houston, TX, 77230, USA. ²⁶ Department of Clinical Science, University of Bergen, 5020 Bergen, Norway. ²⁷ Department of Oncology, Haukeland University Hospital, 5021 Bergen, Norway. ²⁸ Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, New York 10065, USA. ²⁹ Synergie Lyon Cancer, Centre Léon Bérard, 28 rue Laënnec, Cedex 08 Lyon, France. ³⁰ Equipe Erable, INRIA Grenoble-Rhône-Alpes, 655, Av. de l'Europe, 38330 Montbonnot-Saint Martin, France. ³¹ Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital The Norwegian Radiumhospital, 0310, Oslo, Norway. ³² K.G. Jebsen Centre for Breast Cancer Research, Institute for Clinical Medicine, University of Oslo, 0310 Oslo, Norway. ³³ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, Cambridgeshire, UK. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.M. (email: j.martens@erasmusmc.nl).

The recent advance in DNA sequencing technologies has added substantially to the knowledge-base of breast cancer, but also reaffirmed its proverbial heterogeneity^{1–8}. Sequencing DNA to greater depth enables analyses in distinguishing the ‘driver’ mutations—thought to be involved in the tumourigenesis—from the ‘passenger’ mutations—those arising in the lifetime of the cancer cell, without affecting the cancer cell’s fitness. We recently reported on the identification of 12 substitution and 6 rearrangement signatures, as well as 7 consensus patterns therein, using whole genome sequence (WGS) data of 560 cases with primary breast cancer⁹. Numbered according to an earlier scheme², some of the 12 substitution signatures had underlying mechanisms likely explaining the signature: for example, signature 1 was described as ‘age-related’ due to the many C>T mutations in an NCG trinucleotide context (the underlined base is mutated). Similarly, signatures 2 and 13 were related to APOBEC-type mutations (predominant C>T and C>G in a TCN context, respectively), while signature 3 lacked specific features but was strongly associated with inactivated BRCA1 and BRCA2. Signature 5 (with unknown causal mechanism) is present in most cases and is primarily characterized by C>T and T>C mutations. Other signatures were found in few cases (for example, signatures 6, 20 and 26) showing signs of mismatch repair deficiency⁹. It is important to realize that most of the breast cancer cases exhibit a blend of these signatures, possibly obfuscating the resulting effects. The very diverse pattern of mutations found in primary breast cancer was previously studied in depth¹ in a large cohort, integrating (epi)genomic, transcriptomic and proteomic data, but this study mainly focused on the driver genes and reported the resulting effect on its expression in solo or in concert in pathway analyses. The possible effects of the mutational signatures themselves on the transcriptome have been studied with less scrutiny.

In the current study we analyse the transcriptome of 266 cases (191 ER-positive/Her-2 negative and 75 triple-negative) by RNA sequencing with available WGS results⁹ to explore the possible consequences of the substitution signatures at the transcriptome level. We report on clinically and biologically relevant gene expression signatures and associate these with the number and character of signature mutations.

Results

Validation of cohort. RNA-sequencing data of 266 primary breast cancer cases with available WGS data were generated and were first analysed and visualized to verify whether the cohort exhibited the archetypical breast cancer patterns. To this end, the 5,000 most variable transcripts were used to correlate and hierarchically cluster all cases resulting in five clusters (Fig. 1). Cluster 1 contained all cases of the basal-like intrinsic subtype^{6,10} and held 92% of ER-negative cases, while the remaining four subclusters (clusters 2–5) were luminal. Among the subtypes, the total number of substitutions was highest in the basal-likes (Kruskal–Wallis (KW) $P < 0.0001$), while within the luminal cases, the more aggressive luminal B (refs 11,12) compared with luminal A type cancer also showed higher numbers of substitutions (Mann–Whitney U -test (MWU) $P < 0.0001$, Supplementary Fig. 1, and previously reported¹).

Next, we evaluated the reported driver genes (derived from WGS data⁹) with at least five events in this cohort (with the top recurrent events presented in Fig. 1). Amplification of *MYC* ($n = 47$, $\chi^2 P < 0.0001$), of *CCNE1* and *CCND3* (both $n = 7$, $\chi^2 P = 0.0003$), of *PIK3CA* ($n = 9$, $\chi^2 P < 0.0001$) and, as expected, mutations in *TP53* (refs 12–14) ($n = 105$, $\chi^2 P < 0.0001$) were predominant in basal subtype cancers, while mutations in *GATA3* (refs 8,15) ($n = 25$, $\chi^2 P = 0.0006$), and *PIK3CA* ($n = 74$,

$\chi^2 P < 0.0001$) and amplification of *CCND1* ($n = 39$, $\chi^2 P < 0.0001$) were largely restricted to luminal breast cancer, with the latter strongly related to luminal B cancers^{16,17} (33/39 cases, $\chi^2 P < 0.0001$). The lobular cancer driver^{18,19} *CDH1* was mutated in lobular cases (15/21, $\chi^2 P < 0.0001$) and mutations of this type were predominantly observed in luminal A cases (12/21, $\chi^2 P = 0.0013$). In line with their oncogenic role, amplified regions had significantly higher expression for *MYC*, *CCND1*, *ZNF217*, *PIK3CA*, *MDM2*, *CCNE1*, *IGF1R*, *KRAS* and *ERBB2* (MWU all $P < 0.05$, Supplementary Fig. 2). The 8p11 locus was identified as recurrently amplified as well, wherein, for example, *FGFR1*, *ZNF703* and *WHSC1L1* are potential driver candidates. These three genes are indeed significantly higher expressed in the amplified cases (data not shown) precluding isolation of the actual driver from the passenger genes. Tumour suppressors showed lower expression in cases with mutations (*CDH1* and *PTEN*; MWU $P < 0.0001$, Supplementary Fig. 2). For *TP53*, however, the type of mutation mattered (Supplementary Fig. 3) with obvious deleterious variants such as frameshifts ($n = 19$, MWU $P = 0.002$) and nonsense substitutions ($n = 19$, MWU $P < 0.0001$) clearly showing lower expression, while increased expression of *TP53* was found in samples with missense substitutions ($n = 47$, MWU $P < 0.0001$) supportive of a potential tumour promoting role of some of these latter type of *TP53* variants²⁰. The type of mutation was furthermore associated with copy neutral LOH (Fisher’s Exact $P = 0.021$), with cases with missense mutations showing less than expected numbers with copy neutral LOH (17.8 expected, 9 observed), again supportive for a dominant role of missense mutations in the *TP53* gene (Supplementary Fig. 3). Finally, we validated earlier reported¹ mutual exclusivity of *PIK3CA* aberrations, a predominant driver in breast cancer, with any event in *PTEN*, *AKT1* or *AKT2* in all 560 cases. Of the 167 cases with a *PIK3CA* aberration, the majority (91%) lacked any event in the three abovementioned genes ($P = 3.6 \times 10^{-5}$ coMEt exact test²¹, Supplementary Fig. 4) suggesting these are core constituents in this driver pathway in breast cancer.

As indicated, many of these observations were reported earlier, confirming the validity of our cohort for subsequent analyses. However, a few additional observations were noted for known driver genes; for *CCND3*, solely found amplified in basal cases, similar expression levels were observed in all but one sample of the amplified cases compared with diploid cases, implying *CCND3* is likely not the sole driver in this amplified region (Fig. 2, top left). TCGA results ($n = 960$ cases) confirmed our finding but did show a significant increase in expression levels of *CCND3* (ANOVA $P < 0.0001$) in amplified cases supportive of it being an oncogene (Fig. 2, top right). However, with the mean log2 expression level in amplified cases being only 1.2-fold higher compared with diploid cases, this is not the effect one would expect from a biological relevant driver in a particular amplified genomic region. Evaluating genes in the vicinity (~ 700 kb) of *CCND3*, only *USP49* and *BYSL* could be considered as putative drivers instead, showing a more than 3 fold higher expression in the amplified cases in our cohort. In TCGA data, differences were less pronounced, with *BYSL* showing the largest fold-change ($1.5 \times$) (Supplementary Fig. 5a).

Next to the above, in our cohort increased *GATA3* expression was observed in the 25 cases with *GATA3* mutations (MWU $P < 0.0001$, Fig. 2 bottom panel), which may point to a dominant, activating role for mutations in this gene. TCGA results confirmed (ANOVA $P < 0.0001$) this, showing a threefold higher expression found in *GATA3* mutated ($n = 95$) compared with wild-type cases ($n = 864$). The vast majority of mutations were found in exon 5 and 6 (22/25 in our cohort and 90/95 in TCGA), but no significant difference was seen in expression levels between

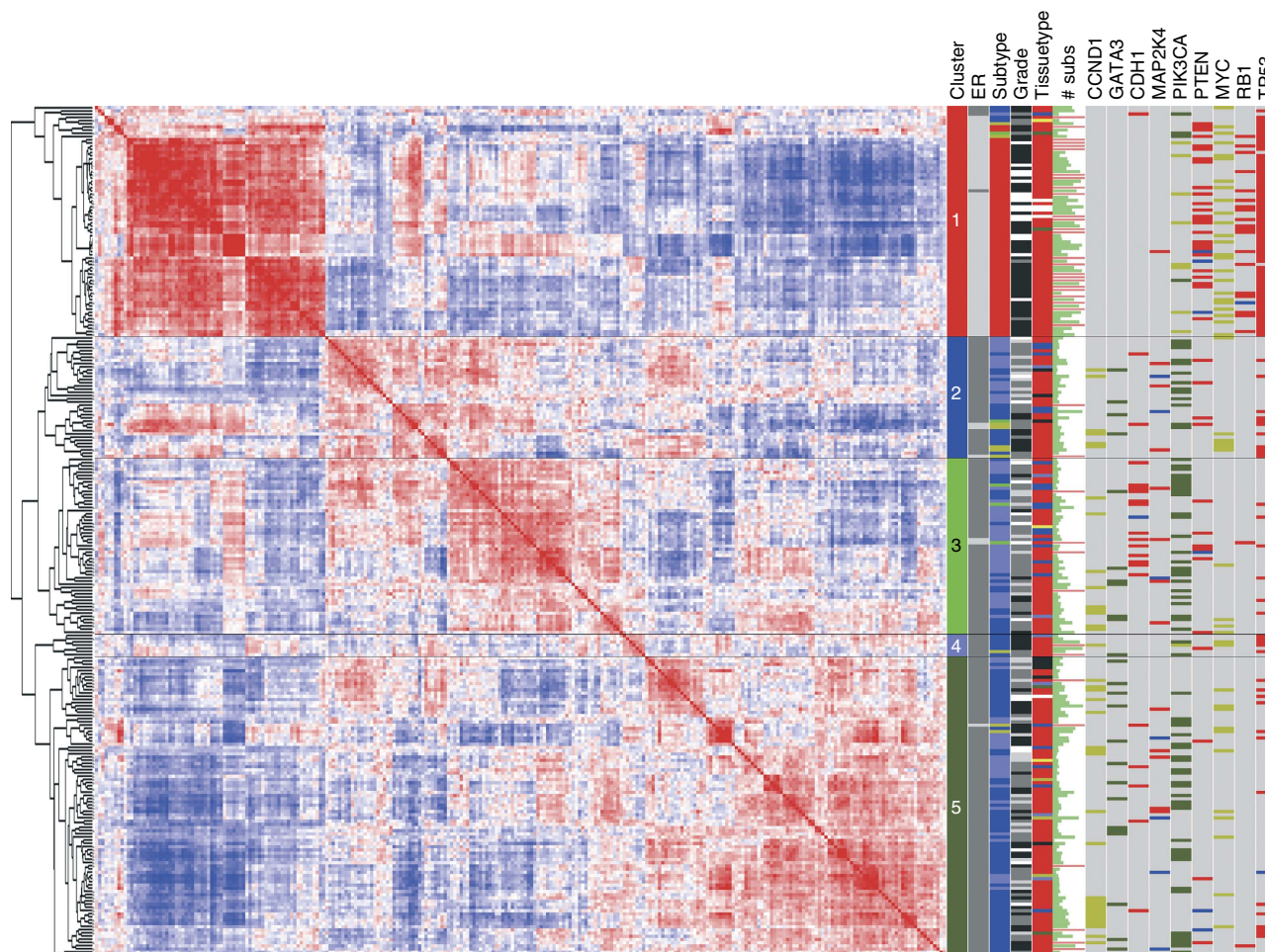


Figure 1 | Clustered correlation matrix of 266 breast cancer cases. The left panel shows the dendrogram and clustered correlation matrix (red is positive, blue negative correlation) of 266 breast cancer cases. The top 5,000 most variable transcripts were used for correlating the samples. For the columns in the right panel, colour codes are as follows: ER: ER-positive dark grey, ER-negative light grey. Subtype: Red, basal; dark blue, Luminal B; light blue, Luminal A; green, normal-like; and dark yellow, her2. Grade: white, NA; light grey, grade 1; grey, grade 2; and black, grade 3. Tissue type: red, ductal; blue, lobular; light blue, micropapillary; grey, mucinous; dark yellow, papillary; yellow, apocrine; and dark green, other type. # subs: The length of the green bar is proportional to the number of substitutions. Cases with >10,000 substitutions are shown with a soft-red coloured bar of equal length. The remaining nine columns show the status of driver genes. Light grey, wild type; dark yellow, copy-number amplification; blue, homozygous deletion; and mutations (substitution, indels, rearrangements) are dark green if activating and red if inactivating.

exon 5 or 6 mutated cases, thus expression changes are independent of the fact that exon 5 mutations predominantly are frameshifts leading to shorter proteins and exon 6 mutations are frameshifts causing proteins with a C-terminal extension (Supplementary Fig. 5b). Noteworthy, the 4% of TCGA cases with an amplification of *GATA3* showed significantly lower expression levels compared with diploid cases (sixfold reduction, ANOVA $P < 0.0001$, data not shown).

Cell cycle and immune pathways and mutational signatures.

Switching from single genes to studying the effects of the various substitution signatures⁹ on the global transcriptome, we first studied the molecular subtypes in relation to the signatures. Significantly higher numbers of signature 3 substitutions but lower number of signature 2 substitutions were found in the basal subtype (KW both $P < 0.0001$) compared with the other subtypes. Substitutions of signature 5 were more abundant in the Her-2 group (KW $P < 0.0001$) while signatures 8 and 13 again showed increased numbers in the basal subtype (KW both $P < 0.0001$). Additional details can be found in Supplementary Fig. 6.

To further evaluate transcriptomic pathways and substitution signatures, we grouped samples according to their dominant mutation signature and compared the cases with, to those lacking that dominant signature (Methods section). The most prominent pathways emerged in ER-positive cases (Supplementary Table 1/ Supplementary Fig. 7) and related to cell cycle, which was significantly associated with the total number of substitutions, while several immune-response pathways were found associated with cases showing predominantly signatures 2 and/or 13. These latter substitution signatures have been accredited⁵ to activity of the APOBEC/AID family of cytidine deaminases, targeting the TCN trinucleotide sequence (the underlined base gets mutated). In ER-negatives, several metabolism related pathways were found related with signature 3 substitutions, but these were predominantly driven by few genes or otherwise were always higher expressed in cases with low numbers of that signature (green bars in the geneplot, Supplementary Fig. 7a), giving less confidence that these pathways were of interest. In contrast, the cell cycle pathways consistently showed multiple genes higher expressed in the ER-positive cases with high numbers of the signature (see Supplementary Fig. 7b,c for examples).

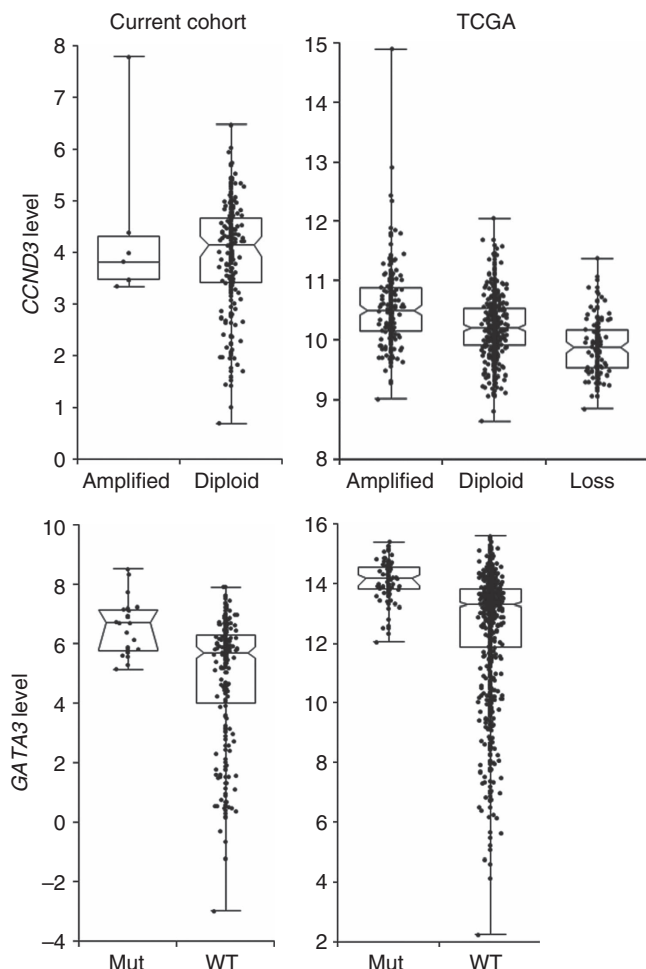


Figure 2 | Boxplots of *CCND3* and *GATA3* expression. All y axes show expression levels in our cohort ($n = 266$, left panels, log2-FPKM) and from TCGA cases ($n = 960$, right panels, RNA Seq V2 RSEM, log2) according to copy-number state of *CCND3* (top panel, our cohort $n = 7$ amplified, TCGA $n = 152$ with loss and $n = 233$ amplified) and *GATA3* mutation state (bottom panel, our cohort $n = 25$ mutated, TCGA $n = 95$ mutated). The box is bounded by the first and third quartile with a horizontal line at the median, whiskers extend to the maximum and minimum value. The notch shows the 95% CI of the median.

To validate the pathway results, we studied cell cycle and immune-related gene expression signatures in further detail.

As an objective readout of the cell cycle phenotype, we used genes annotated to the mitotic cell cycle (MCC, GeneOntology:0000278) and associated global expression of these genes to the number and type of substitutions (signatures 6, 17, 20, 26 and 30 were excluded since very few cases harboured these substitution types) and also to pathological grade (Fig. 3a). Next to confirming that the total number of substitutions were higher in the top quartile of samples with high MCC gene expression (MWU $P < 0.0001$), we found that several signatures did similarly, especially signatures 13, 8, 1 and 3 (MWU $P < 0.0001$, $P = 0.0002$, $P = 0.002$ and $P = 0.005$, respectively). Insofar data were available, pathological grade III was also clearly enriched in the top quartile ($\chi^2 P < 0.0001$). Thus, irrespective of the type of signatures identified in a patient, a higher number of substitutions was related to increased cell proliferation. Adverse outcome is thus to be expected in patients with a high mutational load, and this was confirmed by evaluating overall and relapse-free survival data (Fig. 3b,c).

The other major theme in the analysis of the transcriptome was the involvement of immune response in relation with particularly the APOBEC-type substitution signatures 2 and 13, also in ER-positive cases. Revisiting the pathway results (Supplementary Fig. 7c) showed a clear association of, for example, *IFNG*, *CTLA4* and several chemokines with these 2 APOBEC-signatures, substantiating a relation with T-cell response. To study objectively if indeed activation of the immune system is mutational signature specific, we related in ER-positive cases ($n = 192$) a previously developed tumour-infiltrating lymphocytes (TIL) RNA expression signature for breast cancer²² to the diverse mutational signatures, again excluding signatures 6, 17, 20, 26 and 30 due to low numbers (Fig. 4 and Table 1). Samples with overall high expression of the TIL-signature genes (top quartile of cases) had a significantly higher number of signature 13 (MWU $P = 0.0027$) and also of signature 3 (MWU $P = 0.027$). In contrast, the age-related signature 1 (MWU $P = 0.0005$) displayed a significantly lower number in the high TIL group. The number of mutations of the other signature types was not significantly associated with high TIL expression. Next to analysing absolute numbers, we also analysed the data proportionally, by correcting for the total number of mutations, which removes total mutational load as possible confounding factor but this analysis yielded comparable results (Supplementary Table 2).

Even though the number of patients at risk is low, patients with high TIL and low MCC have a significantly better outcome than those with low TIL and high MCC (Fig. 5a,b). We validated this finding in an independent cohort of 625 lymph-node negative, not (neo)adjuvantly hormonal/chemotherapy treated cases (Fig. 5c). The TIL and MCC groups analysed separately also show significant survival differences in this independent cohort (logrank $P = 0.001$ and $P < 0.0001$ for TIL and MCC, respectively. See Supplementary Fig. 8) and when TIL and MCC groups were analysed in a multivariate model, both remained significant (TIL Hazard Ratio 0.53 (95% confidence interval (CI) 0.36–0.79), $P = 0.002$ and MCC HR 1.81 (95% CI 1.33–2.47), $P < 0.0001$).

Next, we investigated the pathological infiltrate status in all ER-positive cases ($n = 266$, which includes those without RNAseq data). We associated the number of mutations per signature with the pathological infiltrate status. Lymphocytic infiltrate was grouped into three groups, having no, mild and moderate-severe infiltrate, and a test for trend across ordered groups showed significantly higher numbers of signatures 3 and 13 (both $P < 0.0001$, see Table 2) in cases with increasing infiltrate, with signature 2 (the other APOBEC signature) showing a weak association as well (test for trend $P = 0.022$). Of note, the results for signature 3 may be hampered by the fact that many cases lack signature 3 mutations entirely; of the $n = 266$ cases in this analysis 230 (86%) have zero signature 3 mutations. An analysis with the proportion of substitutions of a signature yielded largely similar results (Supplementary Table 3), in the sense that signatures 3 and 13 remain strongly associated with the infiltrate status (test for trend $P < 0.0001$), whereas signature 1 and 5 showed higher proportions in cases with decreasing infiltrate (test for trend $P = 0.015$ and $P < 0.0001$, respectively). To exclude that individual drivers themselves were not responsible for the observed association we investigated if mutation (including substitution, indel, rearrangement, copy-number variation) of the driver genes was directly associated with infiltrate levels. The following genes were analysed: *ARID1A*, *CCND1*, *CDH1*, *GATA3*, *MAP2K4*, *MAP3K1*, *MLL3*, *MYC*, *PIK3CA*, *PTEN*, *TP53* and *ZNF217* ($n = 10, 32, 17, 22, 17, 20, 11, 21, 64, 13, 33$ and 11, respectively), but none of the investigated genes was found to be significantly associated with infiltrate status (Fisher's Exact test, $P > 0.05$).

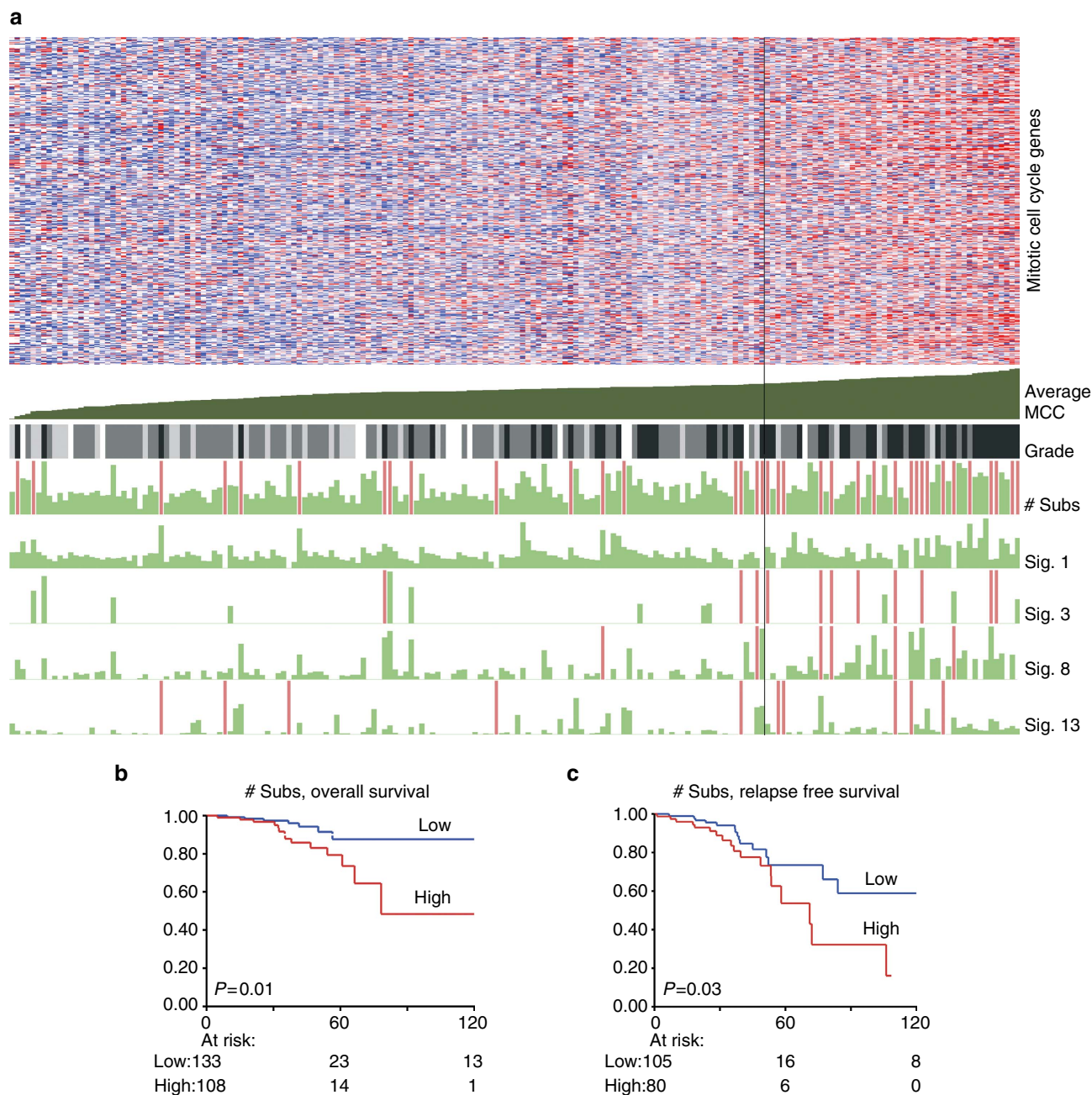


Figure 3 | Mitotic cell cycle gene activity related to mutational signatures and outcome. (a) The average expression of genes ($n=409$) from Gene-Ontology term Mitotic Cell Cycle (GO:0000278) were used to rank ER-positive samples. The vertical black line indicates the third quartile border. Top panel: heatmap of median centred expression values in log2-FPKM, red indicates above median, blue below median expression. Genes are in rows, samples in columns. Below the heatmap: 'Average MCC' shows the average expression of the cell cycle genes. Grade: pathological grade; white, NA; light grey, grade 1; grey, grade 2; and black, grade 3. Last five rows: the length of the green bar is proportional to the number of substitutions. # subs: the total number of substitutions, samples with $>10,000$ substitutions are shown with a soft-red coloured bar and are of equal length. For the columns labelled Sig. 1, 3, 8 and 13, soft-red indicate samples $>3,000$ of such substitutions. (b,c) Overall and relapse-free survival Kaplan-Meier curves. Blue indicates patients with less than the median number of substitutions, red indicates higher than median. P values are logrank-test values. The x-axis shows time in months, y-axis shows the proportion of surviving patients.

Amino acid properties and immune response. To try to understand why only certain mutation signatures relate to a more effective immune response, we determined *in-silico* which substitutions yield predicted neo-epitopes^{23,24} but found no striking signature specific difference in the fraction of predicted neo-epitope presenting substitutions (Supplementary Fig. 9a). We also studied the altered chemical properties^{25,26} of the mutational signature induced amino acid changes. However, changes in

hydrophobicity (Supplementary Fig. 9b) were not strongly differently proportioned across the substitution signatures. In contrast, amino acid substitutions resulting from particularly signature 13 but also from signature 2 displayed a clear increase in charge (Supplementary Fig. 9c), while substitutions caused by signature 1—which were associated with lower infiltrate—displayed a loss of a charge instead. The number of signature 13 substitutions that lead to an increase in charge was

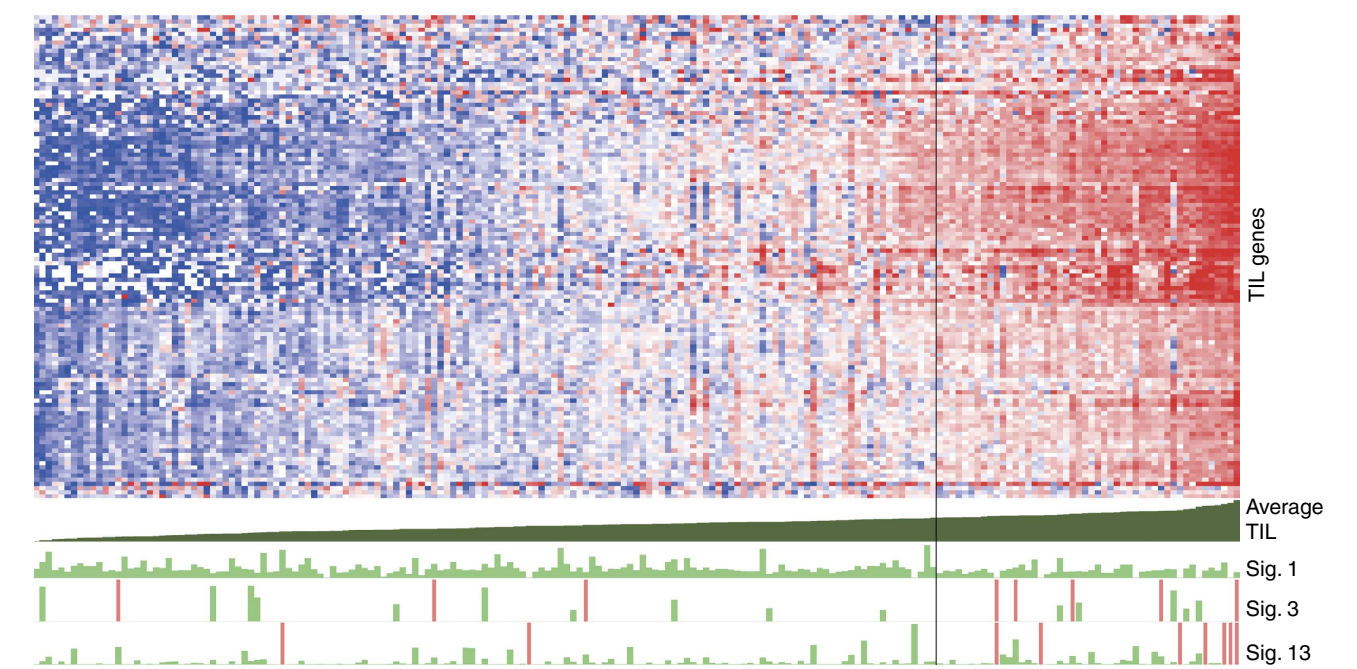


Figure 4 | Activity of TIL-signature genes related to mutational signatures. The average expression of genes ($n = 116$) from a TIL specific RNA-signature was used to rank ER-positive samples. The vertical black line indicates the third quartile border. Top panel: heatmap of median centred expression values in log2-FPKM, red indicates above median, blue below median expression. Genes are in rows, samples in columns. Below the heatmap: the first row shows the average expression of the TIL genes. Last three rows: the length of the green bar is proportional to the number of substitutions of the indicated signatures. Samples with $> 3,000$ substitutions are shown with a soft-red coloured bar and are of equal length.

Table 1 Association of the number of substitutions per signature by TIL-signature group.				
Signature	<i>n</i>	MWU*	Average high TIL	Average low TIL
1	183	0.0005	532.8	752.5
2	179	0.64	2,313.1	328
3	24	0.027	581.4	209.5
5	166	0.27	1,033	1,022.2
6	2	N.D.	17,791	0
8	123	0.59	663.6	453.3
13	115	0.0027	1,897	309.5
17	6	N.D.	9.7	23
18	32	0.39	37.8	64.4
20	1	N.D.	0	0
26	3	N.D.	171.5	87.6
30	1	N.D.	0	0

N denotes the number of samples with > 0 of a particular substitution signature. The average number of substitutions is listed per TIL-signature group; high TIL indicates the top quartile of samples, low TIL the rest of the samples.
*P value of MWU.

significantly associated with increasing lymphocytic infiltrate (test for trend $P < 0.0001$); a similar test for signature 2 substitutions showed a P value of 0.014. Analysing the total number of substitutions leading to increased electric charge, irrespective of signature type, also showed increasing numbers with increasing lymphocytic infiltrate (test for trend $P < 0.0001$) which remained significant even when signature 13 substitutions were excluded (test for trend $P < 0.0001$).

Discussion

In this paper we investigated the mutation signatures present in primary breast cancer and their relation to the transcriptome. Although somewhat impeded by the fact that most tumours display a multiplicity of signatures, we were able to show that

individual signatures have specific effects on the gene expression phenotype of tumour cells. As may be noted from the results, the main observations were obtained in ER-positive cases. The lack of results in ER-negative cases is attributed to the more homogeneous nature of this subtype; for example, virtually all ER-negative patients have high infiltrate levels (in our cohort only four cases display a null infiltrate phenotype) and a high mutational load (just two ER-negative cases are among the bottom 20% after ranking on mutational load), which impedes identifying expression differences due to the lack of a sizeable contrast group.

Regarding the global transcriptome in relation to the substitution signatures, first, we observed that the total number of substitutions was positively associated with expression of cell cycle genes, independent of which of the individual substitution signatures contributed to the total. Reviewing the driver genes, as expected *CCND1* and *MYC* amplified cases, as well as *TP53* mutated cases were enriched in the top quartile MCC expression (Fisher's Exact test, all $P < 0.0001$). The relation between an active cell cycle and various known drivers of cell cycle progression (*CCND1*, *MYC*, *TP53*) and poor outcome in ER-positive patients has been widely investigated^{16,27–31}, thus our observation that the mutational burden (defined here as the total number of substitutions, regardless of type) was also associated with shorter DFS and OS may be not surprising. The association of mutational burden and poor outcome in ER-positive patients was described earlier as well³², but using data based on exome sequencing, which we firmly validated here using whole genome sequencing data, with the latter enabling removal of the possible bias which may exist on the total mutational burden derived from the transcribed part (exome) compared with the entire genome (WGS). Noteworthy is that both mutational signatures which are replication dependent (for example, signature 13) and those that are not (for example, signature 8)⁹ contribute to the mutational burden. This suggests that mutational load as a result of

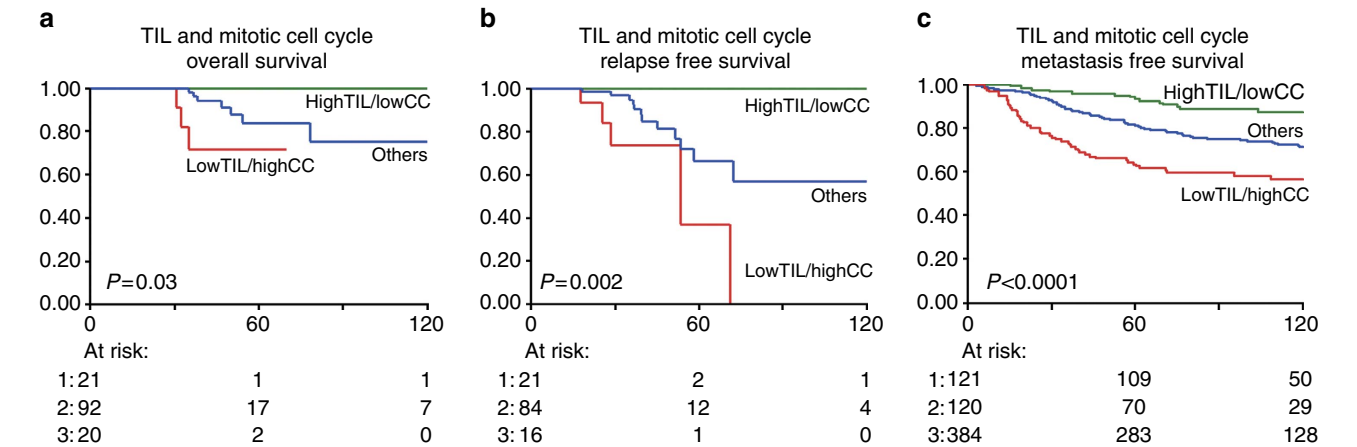


Figure 5 | Combined MCC and TIL-signature genes and outcome. Overall (a) and relapse-free (b) survival Kaplan – Meier curves of our cohort and (c) metastasis-free survival of independent in-house and public data sets. Green line indicates patients with high expression of TIL genes (top quartile) and low expression of MCC genes (bottom three quartiles). The red line indicates patients with low expression of TIL genes and high expression of MCC genes. Blue indicates the remaining patients. *P* values are logrank-test for trend values. The x-axis shows time in months, y-axis shows the proportion of patients. For the numbers at risk, 1 indicates the highTIL/lowCC group (green line), 3 indicates the lowTIL/highCC (red line) and 2 the remaining patients.

Table 2 Association of the number of substitutions per signature by lymphocytic infiltrate group.					
Signature	N	Cuzick*	Average mild	Average moderate/severe	Average moderate/severe
1	253	0.537	592.9	728.5	649.9
2	248	0.022	169.5	657.7	1,401.5
3	36	<0.0001	127.7	197.8	944.2
5	237	0.158	1,131.2	1,040.7	1,059.4
6	2	N.D.	0	54.9	355.8
8	172	0.117	322.4	435	785.4
13	174	<0.0001	91.5	573.1	1,613.9
17	9	N.D.	11.7	20.9	14.5
18	45	0.935	32.6	90.3	34.2
20	0	N.D.	0	0	0
26	3	N.D.	0	126	164.6
30	0	N.D.	0	0	0

N denotes the number of samples with >0 of a particular substitution signature. The average number of substitutions is listed per lymphocytic infiltrate group.
**P* value of Cuzick's nonparametric test for trend across ordered groups.

insufficient or improper repair during DNA replication, as well as those from other sources, both contribute to more rapid progression of ER-positive disease.

Second and interestingly, we observed a significant, positive relation between substitutions of signatures 13 and 3 and a TIL gene expression signature and also to increasing levels of TILs. The association of signature 3 substitutions with immune-response pathways is not readily explained by the type of substitutions—which are broadly patterned—or by the investigated amino acid properties; we speculate that the strong association of signature 3 with inactivated *BRCA1/2* (ref. 9) could make this signature a proxy for BRCA-ness (with its particular rearrangements) in ER-positive cases. Inactivated *BRCA* is tightly linked with loss of homologous recombination (HR) repair mechanisms, leading to distinct rearrangement patterns⁹. We speculate that these rearrangements, which can comprise up to hundreds of events in a sample, leading to randomly aberrant proteins, potentially triggering an immune response. And, although the exact *modus operandi* of how signature 3 or BRCA-ness could increase immune response is as of yet obscure, higher lymphocytic infiltrate have been previously reported in familial *BRCA1* (ref. 33) and *BRCA2*-affected tumours³⁴, the

former study also showing significantly better relapse-free survival for BRCA-patients with higher infiltrate levels.

The connection between TIL and signature 13 substitutions may be based on the distinguishing feature of signature 13; its clear APOBEC pattern^{2,5}, deaminating cytidines in a TCN context, favouring C to G substitutions. We hypothesize that TILs are enriched at sites where the proper peptides are presented; a possible way signature 13 substitutions could activate this immune-response is via its resulting neo-epitopes. One difference we observed for altered peptide sequences resulting from signature 13 substitutions as compared with other equally abundant ones originating from other sources (for example, signature 1, 5, 8 substitutions) was not the number of predicted neo-epitopes, but rather its tendency to being positively charged. The observed increase in charge in signatures 2 and 13 can be readily explained by the fact that these signatures frequently affect the only two negatively charged amino acids Glu and Asp, respectively, which are modified in this mutational context into Lys (signature 2) and His (signature 13), both positively charged. Thus, these latter 2 signatures are most efficient in generating mutated peptides with increased electric charge. Interestingly, patients with an increased lymphocytic infiltrate showed, irrespective of signature type, higher numbers of positively charged amino acid substitutions, indicating that the association of signature 13 with TIL might be because positively charged amino acids are more commonly generated in this signature type.

Concluding, we show that ER-positive breast cancer responds to DNA mutations with two specific changes in the transcriptome. On the one hand, the cell cycle with its associated adverse clinical outcome appears to be more active with an increasing number of mutations, regardless of their underlying signatures. On the other hand, we observe a signature specific association with immune response, possibly arising from an increased number of neo-antigens that consequently may stimulate the immune response more effectively. However, independent confirmation of our finding is needed as well as mechanistic studies explaining the observed association. Still, our results augment earlier reports^{35–37}, which postulate that a high mutational load is sufficient for the immune system to sense one or more neo-epitopes as non-self. However, by untangling the mutational load into specific signatures, our results suggest that the breast tumour cell, while gaining advantage by those mutations stimulating proliferation, may inadvertently provoke

the immune system more effectively if the mutational load is moulded through specific mutational processes. Exploiting this knowledge by purposefully augmenting T-cell reactivity against amino acid substitutions resulting from the proper DNA substitution type could improve cancer immunotherapies.

Methods

Cohorts. The sample cohort used in the study has been described in detail recently⁹. To recapitulate, DNA and RNA of breast cancer tumours were analysed using WGS (WG-DNA data), copy-number analysis, methylation profiling, RNA sequencing and miRNA analysis. Here, we studied all patients with available RNA-seq and WG-DNA data ($n = 266$, no replicates). Sequencing protocols, QC and post-processing of data as well as the DNA substitution and rearrangement signatures and data availability are described elsewhere⁹. For the RNA-sequence cohort, IHC-scored Her2-positivity (as ascertained by a panel of pathologists according to the current standard practice) was an exclusion criterion, though at the time of analysis, some cases showed DNA amplification of the Her2 locus. Gene expression values were available as log2-FPKM values for 49,738 transcripts. For the correlation matrix (Fig. 1) transcripts with >20% missing data were excluded, and the top 5,000 variable transcripts—which included besides 3,484 protein-coding genes, 1,516 non-coding RNA species such as lincRNAs, snRNAs and snoRNAs and so on—were median centred and used to correlate (Pearson) all samples with each other. The resulting correlation matrix was hierarchically clustered³⁸ (Cluster 3.0) using uncentered correlation as distance metric. The molecular subtypes were established using the AIMS method⁶.

Pathway analysis. Since the transcriptome is heavily driven by ER-status, pathway analyses were performed within ER-positive and -negative breast cancer cases separately. In the pathway analysis each signature was analysed separately: the proportion of substitutions of a particular signature was used for grouping samples. Samples which had at least 50% of a particular signature were compared with samples having <20% of such substitutions, where both groups should have at least 15 cases. In ER-positive cases, signature 5 ($n = 54$ >50% proportion versus $n = 32$ <20%) and the APOBEC-signatures (Sig 2 + 13 combined) with $n = 18$ >50% proportion versus $n = 45$ <20%) were available for analysis and in ER-negative cases, signature 3 ($n = 24$ >50% versus $n = 22$ <20%) could be evaluated. In addition, disregarding the type of signature, the total number of substitutions was used to group samples with the most extreme number of substitutions (top and bottom 20%). Initially samples were grouped in a low and high mutational burden group, separately for ER-positive and -negative cases. However, in the ER-negatives, samples in the bottom 20% still had a high absolute number of substitutions. Since our aim was to identify differential pathways associated with the absolute tumour burden, we wanted to have a bottom 20% group with absolute low numbers of substitutions. Samples were thus ranked according to mutational burden independent of ER-status to obtain the bottom and top 20% groups. In ER-positive cases, this gave a group of $n = 15$ in the top 20% versus $n = 52$ in the bottom 20%. In ER-negative cases there were only two samples in the bottom 20%, precluding a meaningful analysis. Pathway analysis was performed using the R-package 'global test'³⁹ using Biocarta (www.biocarta.com) and KEGG (ref. 40) as annotation databases. All P values were corrected for multiple testing (Bonferroni–Holm) and checked by re-sampling 1,000 times to ascertain the number of times an equally sized, randomly chosen group of genes is at least as significant as the true set of genes belonging to a pathway. Pathways were considered of interest if the P value of the global test after correcting for multiple testing and the re-sampling P value were both below 0.05. Second, pathways were considered relevant pathways if these consistently showed multiple genes higher expressed in the cases with high numbers of a signature.

Neo-antigen prediction and hydrophobicity of amino acids. To ascertain which of the amino acid changing substitutions potentially contain a neo-epitope, all unique, non-silent, non-terminating amino acid substitutions were selected ($n = 26,475$) and we focused our analyses on CD8 T-cell epitopes. FASTA files of the wild-type sequences were obtained via the consensus coding sequence database (<http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi>) and custom Perl scripts were used to mutate the wild-type sequence according to the identified substitution in our cohort. Since the software to evaluate neo-epitopes uses all possible 9-mer peptides in a sequence, we selected a 17-mer sequence with the mutated amino acid in the middle – thus making sure the mutated amino acid was always present in all possible 9mers of that sequence. Care was taken if the substitution was too near to the start or end of a gene to ensure a 9mer was always present. The mutated FASTA sequences were uploaded to NETMHC v3.4 (refs 23–24) (<http://www.cbs.dtu.dk/services/NetMHC/>) to predict the binding affinity to the following HLA-A2 alleles, representing the most prevalent MHC class I alleles among the Caucasian population: HLA-A2:01, 02, 03, 06, 11, 12, 16, 17, 19 and 50. With a possible eight sequences to be tested per gene, this gives a matrix of at most 80 predictions. As defined²³, a predicted EC50 <50 nM of a peptide towards HLA-A2 alleles indicates a 'strong binder' and if any of the possible predictions was found <50 nM the gene was considered harbouring a potential neo-antigen.

Changes in hydrophobicity of mutated amino acids was ascertained using the scale of Kyte & Doolittle⁴¹ where the level of the mutated amino acid was subtracted from the wild-type amino acid. A positive difference (>0) was labelled as 'increase' and <0 as 'decrease' in hydrophobicity. Changes in electric charge were evaluated by assigning Aspartic acid and Glutamic acid as negatively charged and Lysine, Arginine and Histidine as positive. Next, base substitutions were evaluated if a resulting amino acid change led to a change in electrical charge. Since all substitutions are assigned to a mutational signature, the total number of charge changing (and hydrophobicity changing) substitutions per signature was established.

Independent validation set. In-house and publicly available gene expression data of lymph-node negative primary breast cancer patients, who were not treated (neo)adjuvantly with hormonal/chemotherapy and available metastasis-free survival data were used, leading to a cohort of 867 patients ($n = 625$ ER-positive). This selection enables a pure study on prognosis. Data were gathered from Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) entries GSE2034, GSE5327, GSE2990, GSE7390 and GSE11121, with all data available on Affymetrix U133A chip. Raw.cel files were processed using fRMA (ref. 42) parameters (median polish), after which batch effects were corrected using ComBat (ref. 43). The average expression of TIL genes²² and MCC genes (Gene Ontology:0000278) were calculated of each sample, and the samples in the top quartile for TIL and the top quartile for MCC were used to group the patients in three groups; those with high TIL/low MCC, those with low TIL/high MCC and the remaining patients.

Statistics. Associations between gene expression levels and gene-status (mutant, amplification and so on) were analysed using MWU, or KW when more than two categories were evaluated. In the analyses where the expression level of a gene was associated with its own status and not with other genes, no multiple testing correction was necessary. To test for a trend across ordered groups (for example, number of substitutions in increasing grade), Cuzick's nonparametric test for trend was used, an adjunct to the KW test. Pearson's χ^2 test was used to evaluate categories (for example, mutation status versus molecular subtype) or by Fisher's Exact test when number of events were low. Stata v13 (StataCorp, Texas, USA) was used for the statistical tests and Kaplan–Meier survival curves. A two-sided P value of <0.05 was considered statistically significant.

Data availability. WGS, RNA-seq and methylation data that support the findings in this study are available at the European-Genome Phenome Archive (<https://www.ebi.ac.uk/ega>) under the accession code EGAS00001001178. For the validation data, in-house GSE2034 and GSE5327, and otherwise public data GSE2990, GSE7390 and GSE11121 are available at the Gene Expression Omnibus website (<http://www.ncbi.nlm.nih.gov/geo/>). All other data are contained within the Article and its Supplementary files or available from the authors on request.

References

1. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
2. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
3. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
4. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
5. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
6. Paquet, E. R. & Hallett, M. T. Absolute assignment of breast cancer intrinsic molecular subtype. *J. Natl Cancer Inst.* **107**, 357 (2015).
7. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
8. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
9. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
10. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
11. Smid, M. *et al.* Subtypes of breast cancer show preferential site of relapse. *Cancer Res.* **68**, 3108–3114 (2008).
12. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
13. Langerød, A. *et al.* TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast. Cancer Res.* **9**, R30 (2007).
14. Manie, E. *et al.* High frequency of TP53 mutation in BRCA1 and sporadic basal-like carcinomas but not in BRCA1 luminal breast tumors. *Cancer Res.* **69**, 663–671 (2009).

15. Usary, J. *et al.* Mutation of GATA3 in human breast tumors. *Oncogene* **23**, 7669–7678 (2004).
16. Holm, K. *et al.* Characterisation of amplification patterns and target genes at chromosome 11q13 in CCND1-amplified sporadic and familial breast tumours. *Breast. Cancer Res. Treat.* **133**, 583–594 (2012).
17. Roy, P. G. *et al.* High CCND1 amplification identifies a group of poor prognosis women with estrogen receptor positive breast cancer. *Int. J. Cancer* **127**, 355–360 (2010).
18. Berx, G. *et al.* E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers. *EMBO J.* **14**, 6107–6115 (1995).
19. Berx, G. *et al.* E-cadherin is inactivated in a majority of invasive human lobular breast cancers by truncation mutations throughout its extracellular domain. *Oncogene* **13**, 1919–1925 (1996).
20. Walerych, D., Napoli, M., Collavin, L. & Del Sal, G. The rebel angel: mutant p53 as the driving oncogene in breast cancer. *Carcinogenesis* **33**, 2007–2017 (2012).
21. Leiserson, M. D., Wu, H. T., Vandin, F. & Raphael, B. J. CoMet: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome. Biol.* **16**, 160 (2015).
22. Massink, M. P., Kooi, I. E., Martens, J. W., Waisfisz, Q. & Meijers-Heijboer, H. Genomic profiling of CHEK2*1100delC-mutated breast carcinomas. *BMC Cancer* **15**, 877 (2015).
23. Lundegaard, C. *et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res.* **36**, W509–W512 (2008).
24. Nielsen, M. *et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* **12**, 1007–1017 (2003).
25. Chowell, D. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl Acad. Sci. USA* **112**, E1754–E1762 (2015).
26. Kim, Y., Sidney, J., Pinilla, C., Sette, A. & Peters, B. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinformatics* **10**, 394 (2009).
27. Reis-Filho, J. S. & Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* **378**, 1812–1823 (2011).
28. Yu, J. X. *et al.* Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer* **7**, 182 (2007).
29. Berns, E. M. *et al.* c-myc amplification is a better prognostic factor than HER2/neu amplification in primary breast cancer. *Cancer Res.* **52**, 1107–1113 (1992).
30. Liao, D. J. & Dickson, R. B. c-Myc in breast cancer. *Endocr. Relat. Cancer* **7**, 143–164 (2000).
31. Olivier, M. *et al.* The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. *Clin. Cancer. Res.* **12**, 1157–1167 (2006).
32. Haricharan, S., Bainbridge, M. N., Scheet, P. & Brown, P. H. Somatic mutation load of estrogen receptor-positive breast tumors predicts overall survival: an analysis of genome sequence data. *Breast Cancer Res. Treat.* **146**, 211–220 (2014).
33. van Verschuier, V. M. *et al.* Tumor-associated inflammation as a potential prognostic tool in BRCA1/2-associated breast cancer. *Hum. Pathol.* **46**, 182–190 (2015).
34. Bane, A. L. *et al.* BRCA2 mutation-associated breast cancers exhibit a distinguishing phenotype based on morphology and molecular profiles from tissue microarrays. *Am. J. Surg. Pathol.* **31**, 121–128 (2007).
35. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).
36. Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
37. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).
38. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
39. Goeman, J. J., van de Geer, S. A., de Kort, F. & van Houwelingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93–99 (2004).
40. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
41. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
42. McCall, M. N., Bolstad, B. M. & Irizarry, R. A. Frozen robust multiarray analysis (fRMA). *Biostatistics* **11**, 242–253 (2010).
43. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

Acknowledgements

This work has been funded through the ICGC Breast Cancer Working group by the Breast Cancer Somatic Genetics Study (a European research project funded by the European Community's Seventh Framework Programme (FP7/2010–2014) under the grant agreement number 242006); the Triple Negative project funded by the Wellcome Trust (grant reference 077012/Z/05/Z).

Personally funded by grants: FGR-G, SM, KR, SM were funded by BASIS. JAF was funded through an ERC Advanced Grant (ERC-2012-AdG-322737). SN-Z is a Wellcome Beit Fellow and personally funded by a Wellcome Trust Intermediate Fellowship (WT100183MA). ALR is partially supported by the Dana-Farber/Harvard Cancer Center SPOR in Breast Cancer (NIH/NCI 5 P50 CA168504-02). AS was supported by Cancer Genomics Netherlands (CGC.nl) through a grant from the Netherlands Organisation of Scientific research (NWO). MS was supported by the EU-FP7-DDR response project. CD was supported by a grant from the Breast Cancer Research Foundation. EB was funded by EMBL. JE was funded by The Icelandic Centre for Research (RANNIS).

For technical or administrative support: Lira Mamanova from the Wellcome Trust Sanger Institute is thanked for her expertise in RNA library preparation. We thank contributors Luc Dirix (Antwerp, Belgium) and Suet-Feung Chin (Cambridge, UK), Jon G Jonasson (University of Iceland), The Icelandic Cancer Registry and from Nijmegen (The Netherlands): Margrete Schlooz-Vries, Jolien Tol, Hanneke van Laarhoven, Fred Sweep and Peter Bult. Benita Kiat Tee Tan from the National Cancer Centre Singapore and Singapore General Hospital is thanked for contributing patient samples for this study as are OSBREAC, the Oslo Breast Cancer Research Consortium, Norway (<http://ous-research.no/home/kgjebsen/home/14105>), the Brisbane Breast Bank, Australia and the Tayside Tissue Bank, Dundee, UK.

Finally, we acknowledge all members of the ICGC Breast Cancer Working Group.

Author contributions

M. Smid and J.W.M.M. wrote the main paper. M. Stratton, S.M., S.N.-Z., H.G.S., J.A.F., J.W.M.M. were involved in the strategy and supervision of the project. Experiments were performed by F.G.R.-G., A.M.S., W.J.C.P.-V.d.S., M.v.d.V.-D., A.v.G., J.S., A.B.B., K.B., M.R. and A.L. M. Smid, S.N.-Z., J.S., A.B.B., M.J.v.d.V., A.L.R., A.B., H.D.D., R.D., M.E.M.-v.G., C.H.M.v.D., S.R.L., M.R., A.V., A.B.-D., E.B. and J.W.M.M. analysed data. Samples were contributed by J.A.F., J.W.M.M., A.L.R., A.F., C.P., A.M.T., C.C., P.N.S., P.T.S., S.R.L., S.V.L., C.D., S.T., O.A.S., A.B., A.V.-S., P.A.F., S.K., T.K., G.T., A.V., A.L., A.-L.B.-D., G.M., R.S., G.G.G.M.V.d.E.

Additional information

Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunications>

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

How to cite this article: Smid, M. *et al.* Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. *Nat. Commun.* **7**:12910 doi: 10.1038/ncomms12910 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016